To pool or not to pool: a partially heterogeneous framework

Vasilis Sarafidis and Neville Weber

The University of Sydney, Monash University

8. December 2009

A Partially Heterogeneous Framework for Analyzing Panel Data*

Vasilis Sarafidis† University of Sydney

Neville Weber‡ University of Sydney

This version: May 2011

Abstract

This paper proposes a partially heterogeneous framework for the analysis of panel data with fixed T. In particular, the population of cross-sectional units is grouped into clusters, such that slope parameter homogeneity is maintained only within clusters. Our method assumes no a priori information about the number of clusters and cluster membership and relies on the data instead. The unknown number of clusters and the corresponding partition are determined based on the concept of 'partitional clustering', using an information-based criterion. It is shown that this is strongly consistent, i.e. it selects the true number of clusters with probability one as $N \to \infty$. Simulation experiments show that the proposed criterion performs well even with moderate N and the resulting parameter estimates are close to the true values. We apply the method in a panel data set of commercial banks in the US and we find five clusters, with significant differences in the slope parameters across clusters.

Key Words: partial heterogeneity, partitional clustering, exploratory data analysis, information-based criterion, model selection.

JEL Classification: C13; C33; C51.

*We are grateful to Genliang Guan for excellent research assistance. We have also benefited from helpful comments by Geert Dhaene, Daniel Oron, Tom Wansbeek, Yuehua Wu and seminar participants at the Erasmus University Rotterdam, University of Leuven, University of York and the Tinbergen Institute. Financial support from the Research Unit of the Faculty of Economics and Business at University of Sydney is gratefully acknowledged.

†Corresponding author. Faculty of Economics and Business, University of Sydney, NSW 2006, Australia. E-mail: vasilis.sarafidis@sydney.edu.au.

‡School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia. E-mail: neville.weber@sydney.edu.au.

1 Introduction

Slope parameter homogeneity is often an assumption that is difficult to justify in panel data models, both on theoretical grounds and from a practical point of view. On the other hand, the alternative of imposing no structure on how these coefficients may vary across individual units may be rather extreme. This argument is in line with evidence provided by a substantial body of applied work. For example, Baltagi and Griffin (1997) reject the hypothesis of coefficient homogeneity in a panel of gasoline demand regressions across the OECD countries, and Burnside (1996) rejects the hypothesis of homogeneous production function parameters in a panel of US manufacturing industries. Even so, both studies show that fully heterogeneous models lead to very imprecise estimates of the parameters, which in some cases have even the wrong sign. Baltagi and Griffin notice that this is the case despite the fact that there is a relatively long time series in the panel – to the extent that the traditional pooled estimators are superior in terms of root mean square error and forecasting performance. Furthermore, Burnside suggests that in general his estimates show significant differences between the homogeneous and the heterogeneous models and the conclusions about the degree of returns to scale in the manufacturing industry would heavily depend on which one of these two models is used. Along the same line Baltagi, Griffin and Xiong (2000) place the debate between homogeneous versus heterogeneous panel estimators in the context of cigarette demand and conclude that even with a relatively large number of time series observations, heterogeneous models for individual states tend to produce implausible estimates with inferior forecasting properties, despite the fact that parameter homogeneity is soundly rejected by the data. As pointed out by Browing and Carro (2007), there is usually a lot more heterogeneity than what empirical researchers allow for in econometric modelling, although the level of heterogeneity and how one allows for it can make a large difference for outcomes of interest.

These findings indicate that the modelling framework of slope parameter homogeneity (pooling) and full heterogeneity may be polar cases, and other intermediate cases may often provide more realistic solutions in practice. The pooled mean group estimator (PMGE) proposed by Pesaran, Shin and Smith (1999) bridges the gap between pooled and fully heterogeneous estimators by imposing partially heterogeneous restrictions related to the time dimension of the panel. In particular, this intermediate estimator allows the short-run parameters of the model to

be individual-specific and restricts the long-run coefficients to be the same across individuals for reasons attributed to budget constraints, arbitrage conditions and common technologies.

In this paper we propose a modelling framework that imposes partially heterogeneous restrictions not with respect to the time dimension of the panel, as PMGE does, but with respect to the cross-sectional dimension, N. In particular, the population of cross-sectional units is grouped into distinct clusters, such that within each cluster the slope parameters are homogeneous and all intra-cluster heterogeneity is attributed to a function of unobserved individual-specific and/or time-specific effects. The clusters themselves are heterogeneous, that is, the slope parameters vary across clusters.

Naturally, the practical issue of how to group the individuals into clusters is central in the paper. If there is a priori information about cluster membership and the number of clusters, the problem reduces to a split-sample standard panel data regression. In many cases, while it might be plausible to think of a set of factors to which slope parameter heterogeneity can be attributed, such as differences in tastes, beliefs, abilities, skills or constraints, these are often unobserved and moreover provide no guidance as to what the appropriate partitioning is, or how many clusters exist. In addition, there are often several ways to partition the sample and while the formed clusters may be economically meaningful, they may not be optimal from a statistical point of view.

Clustering methods have already been advocated in the econometric panel data literature by some researchers; for instance, Durlauf and Johnson (1995) propose clustering the individuals using regression tree analysis, and Vahid (1999) suggests a classification algorithm based on a measure of complexity using the principles of minimum description length and minimum message length, which are often employed in coding theory.[1] Both these methods are based on the concept of *hierarchical clustering*, which involves building a 'hierarchy' from the individual units by progressively merging them into larger clusters. The proposed algorithms provide a consistent estimate of the true number of clusters for $T \to \infty$ only. On the contrary, this paper proposes estimating the unknown number of clusters and the corresponding partition based on the concept of *partitional clustering*. In

[1] Kapetanios (2006) proposes an information criterion, based on simulated annealing, to address a related problem – in particular, how to decompose a set of series into a set of poolable series for which there is evidence of a common parameter subvector and a set of series for which there is no such evidence.

particular, the underlying structure is recovered from the data by grouping the individuals into a fixed number of clusters using an initial partition, and then re-allocating each individual into the remaining clusters until the final preferred partition minimises an objective function. In this paper the residual sum of squares (RSS) of the estimated model is used as the objective function. The number of clusters is determined by the clustering solution that minimises RSS subject to a penalty function that is strictly increasing in the number of clusters. Intuitively our procedure is identical to a standard model selection criterion method, although the study of the asymptotics is more complicated because the number of individuals contained in a given cluster may vary with N. It is shown that the proposed criterion is strongly consistent, i.e. it estimates the true number of clusters with probability one as N grows, for any T fixed. This is important because most frequently panel data sets entail a large number of individuals and a small number of time series observations. Furthermore, it is usually the case of small T where some kind of pooling provides substantial efficiency gains over full heterogeneity.

As with other clustering procedures, our method relies on the data to suggest any clustering structure that might exist, and as such it can be described as an exploratory data analysis approach. Hence, it can be particularly useful when there is no a priori information about the clustering structure, or when one is interested in examining how far a structure that might be meaningful according to some economic measure lies from the structure that fits the data best.

The remainder of the paper is as follows. The next section formulates the problem. Section 3 analyses the properties of the proposed clustering criterion. Section 4 discusses the implementation of the algorithm used to implement the clustering procedure. The finite-sample performance of the algorithm is investigated in Section 5 using simulated data. Section 6 applies our partially heterogeneous framework to a random panel of 551 banking institutions operating in the US, each observed over a period of 15 years. Five clusters are found and the results show some large and statistically significant differences in the value of the slope coefficients across these clusters. A final section provides concluding remarks.

2 Model Specification and Cluster Determination

We consider the following panel data model:

$$y_{\omega it} = \boldsymbol{\beta}'_{\omega}\mathbf{x}_{\omega it} + u_{\omega it}, \tag{1}$$

where $y_{\omega it}$ denotes the observation on the dependent variable for the i^{th} individual that belongs to cluster ω at time t, $\boldsymbol{\beta}_{\omega} = (\beta_{\omega 1}, ..., \beta_{\omega K})'$ is a $K \times 1$ vector of fixed coefficients, $\mathbf{x}_{\omega it} = (x_{\omega it1}, ..., x_{\omega itK})'$ is a $K \times 1$ vector of covariates, and $u_{\omega it}$ is a disturbance term. Therefore, each cluster has its own regression structure with $\omega = 1, ..., \Omega_0$, $i\,[\in \omega] = 1, ..., N_{\omega}$, and $t = 1, ..., T$. This means that the total number of clusters equals Ω_0, the ω^{th} cluster has N_{ω} individuals, for which there are T time series observations available. The total number of individuals in all clusters equals $N = \sum_{\omega=1}^{\Omega_0} N_{\omega}$ and the total sample size is given by $S = NT$.

If the true number of clusters and the corresponding partition or membership of individual i into cluster ω are both known, the problem reduces to a split-sample standard panel data regression, which is straightforward enough to estimate. In this paper we are interested in estimating the vector $\boldsymbol{\beta}_{\omega}$ for $\omega = 1, ..., \Omega_0$, when neither the true number of clusters nor cluster membership are known. Unfortunately, ignoring cluster-specific slope parameter heterogeneity by pooling the data will not provide a consistent estimate of $\boldsymbol{\beta} = \sum_{\omega=1}^{\Omega_0} \frac{N_{\omega}}{N}\boldsymbol{\beta}_{\omega}$, which is the natural weighted average value of the cluster-specific coefficients with weights determined by the proportion of individuals belonging to each cluster. This holds true even under strict exogeneity of the regressors.

To see this, let $E\left(u_{\omega it}|\mathbf{x}_{\omega i1}, ..., \mathbf{x}_{\omega iT}\right) = \mathbf{0}$ and $N_{\omega}^{-1}\sum_{i=1}^{N_{\omega}} X'_{\omega i}X_{\omega i} \xrightarrow{p} M_{XX,\omega}$, a finite and positive definite matrix, where $X_{\omega i} = (\mathbf{x}_{\omega i1}, ..., \mathbf{x}_{\omega iT})'$. The pooled least-squares estimate of $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{pooled} &= \left[\sum_{\omega=1}^{\Omega_0}\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega_0}\sum_{i=1}^{N_\omega} X'_{\omega i}\mathbf{y}_{\omega i}\right] \\
&= \left[\sum_{\omega=1}^{\Omega_0}\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i}\right]^{-1}\sum_{\omega=1}^{\Omega_0}\left(\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i}\right)\left(\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i}\right)^{-1}\left(\sum_{i=1}^{N_\omega} X'_{\omega i}\mathbf{y}_{\omega i}\right) \\
&= \sum_{\omega=1}^{\Omega_0}\left[\sum_{\omega=1}^{\Omega_0}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i}\right)\right]^{-1}\left[\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i}\right)\right]\widehat{\boldsymbol{\beta}}_\omega \\
&= \sum_{\omega=1}^{\Omega_0}\widehat{W}_\omega\widehat{\boldsymbol{\beta}}_\omega. \qquad (2)
\end{aligned}
$$

The expression above shows that $\widehat{\boldsymbol{\beta}}_{pooled}$ is a matrix-weighted average of the cluster-specific estimates, where the weights are inversely proportional to the cluster covariance matrices. Therefore, letting $N_\omega/N \to c_\omega$, $\widehat{\boldsymbol{\beta}}_{pooled}$ converges in probability to

$$
\widehat{\boldsymbol{\beta}}_{pooled} \xrightarrow{p} \sum_{\omega=1}^{\Omega_0} W_\omega\boldsymbol{\beta}_\omega, \qquad (3)
$$

where $W_\omega = \left[\sum_{\omega=1}^{\Omega_0} c_\omega M_{XX,\omega}\right]^{-1}[c_\omega M_{XX,\omega}]$. The pooled least-squares estimator is not consistent for $\boldsymbol{\beta}$ unless, say, the limiting matrix $M_{XX,\omega}$ is constant across clusters. The condition $M_{XX,\omega} = M_{XX}$ is unnatural in economic data sets and therefore it is unlikely to hold true in most empirical applications.[2]

Our aim is to try to determine whether a clustering structure can be identified among individuals without utilising a priori information, but rather by relying on the data to suggest any possible groups. Let us denote the true partition of the N individuals into Ω_0 clusters by $\Pi_{\Omega_0} = \{C^0_1, \ldots, C^0_{\Omega_0}\}$, where C^0_ω is a set of indices of elements in the ω^{th} cluster such that $C^0_\omega = \{\omega_1, \ldots, \omega_{N_{0\omega}}\} \subseteq \{1, 2, \ldots, N\}$. Thus, the number of individuals in the ω^{th} cluster is $|C^0_\omega| = N_{0\omega}$, and $N_{01} + \ldots + N_{0\Omega_0} = N$.

The model under the true partition will be expressed as follows:

$$
y_{C^0_{\omega_{it}}} = \boldsymbol{\beta}'_{0\omega}\mathbf{x}_{C^0_{\omega_{it}}} + u_{C^0_{\omega_{it}}},\; u_{C^0_{\omega_{it}}} = \boldsymbol{\lambda}'_{C^0_{\omega_{it}}}\boldsymbol{\phi}_t + \varepsilon_{C^0_{\omega_{it}}}, \qquad (4)
$$

[2]Fernandez-Val (2005) and Graham and Powel (2009) study the estimands of linear panel fixed effects estimators in random coefficient models. Yitzhaki (1996) investigates the properties of pooled OLS in a cross-sectional model with individual-specific coefficients. In these studies the underlying assumption is that the data generating process takes a correlated random coefficients form, i.e. the individual-specific parameters are correlated with the regressors. In contrast, our result holds even under strict exogeneity of the covariates.

or, in matrix form,

$$\mathbf{Y}_{C_\omega^0} = X_{C_\omega^0}\boldsymbol{\beta}_{0\omega} + u_{C_\omega^0},\ u_{C_\omega^0} = (I_{N_{0\omega}} \otimes \Phi)\,\boldsymbol{\lambda}_{C_\omega^0} + \varepsilon_{C_\omega^0}, \tag{5}$$

where $\mathbf{Y}_{C_\omega^0} = \left(\mathbf{y}'_{\omega_1}, ..., \mathbf{y}'_{\omega_{N_{0\omega}}}\right)'$, with $\mathbf{y}_{\omega_i} = (y_{\omega i1}, ..., y_{\omega iT})'$, is the $(N_{0\omega}T)\times 1$ vector of observations on the dependent variable for the individuals in the ω^{th} cluster, $X_{C_\omega^0} = \left(\mathbf{x}'_{\omega_1}, ..., \mathbf{x}'_{\omega_{N_{0\omega}}}\right)'$, with $\mathbf{x}_{\omega_i} = (\mathbf{x}_{\omega i1}, ..., \mathbf{x}_{\omega iT})'$, is the $(N_{0\omega}T) \times K$ matrix of covariates and $\boldsymbol{\beta}_{0\omega}$ is a vector of fixed coefficients specific to each cluster.

The error term is subject to a factor structure where $\Phi = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_T)'$ is a $T \times r$ matrix of unobserved common factors and $\boldsymbol{\lambda}_{C_\omega^0} = \left(\boldsymbol{\lambda}'_{\omega_1}, ..., \boldsymbol{\lambda}'_{\omega_{N_{0\omega}}}\right)'$ is a $N_{0\omega}r \times 1$ vector of factor loadings. Thus, the error allows for individual-specific unobserved heterogeneity, captured by $\boldsymbol{\lambda}_{\omega_i}$, that varies over time in an intertemporally arbitrary way, albeit in a similar fashion across i. It also allows for the presence of common unobserved shocks (such as technological shocks and financial crises), captured by $\boldsymbol{\phi}_t$, the impact of which is different for each individual i. Both cases can be thought of as generating cross-sectional dependence. The composite error term reduces to the usual two-way error components model by setting $r = 2$, $\phi_t = (1, \gamma_t)'$ and $\boldsymbol{\lambda}_{\omega_i} = (\eta_i, 1)'$. The unobserved factors, $\boldsymbol{\phi}_t$, could be correlated with $\mathbf{x}_{\omega it}$ and to allow for such a possibility the following specification for the covariates will be considered:

$$X_{C_\omega^0} = (I_{N_{0\omega}} \otimes \Phi)\,\Lambda_{C_\omega^0} + V_{C_\omega^0}, \tag{6}$$

where $\Lambda_{C_\omega^0}$ is a $N_{0\omega}r \times K$ matrix of factor loadings and $V_{C_\omega^0}$ is a $(N_{0\omega}T) \times K$ matrix containing the idiosyncratic errors of the covariates, which are distributed independently of the common effects and across i.

Pre-multiplying (5) by the transformation matrix $Q_{C_\omega^0} = I_{N_{0\omega}T} - I_{N_{0\omega}} \otimes \left[\Phi\left(\Phi'\Phi\right)^{-1}\Phi'\right]$ that eliminates the factor structure yields

$$Q_{C_\omega^0}\mathbf{Y}_{C_\omega^0} = Q_{C_\omega^0}X_{C_\omega^0}\boldsymbol{\beta}_{0\omega} + Q_{C_\omega^0}\boldsymbol{\varepsilon}_{C_\omega^0}, \tag{7}$$

or

$$\widetilde{\mathbf{Y}}_{C_\omega^0} = \widetilde{X}_{C_\omega^0}\boldsymbol{\beta}_{0\omega} + \widetilde{\boldsymbol{\varepsilon}}_{C_\omega^0}, \tag{8}$$

where $\widetilde{\mathbf{Y}}_{C_\omega^0} = Q_{C_\omega^0}\mathbf{Y}_{C_\omega^0}$, $\widetilde{X}_{C_\omega^0} = Q_{C_\omega^0}V_{C_\omega^0}$, and so on.

Suppose we partition the population into Ω clusters $\Pi_\Omega^{(N)} = \left\{C_{\Omega 1}^{(N)}, ..., C_{\Omega\Omega}^{(N)}\right\}$ and assume the true number of clusters is bounded by some constant ξ. For ease of notation we will drop the (N) superscript unless there is ambiguity. Let $\widehat{\boldsymbol{\beta}}_{\Omega\omega}$ be the least squares estimate of $\boldsymbol{\beta}$ based on the observations in cluster $C_{\Omega\omega}$, $\widehat{\boldsymbol{\beta}}_{0\omega}$

be the least squares estimate of $\boldsymbol{\beta}$ based on the observations in the true cluster C_ω^0 and $\widehat{\boldsymbol{\beta}}_{\omega|j}$ be the least squares estimate based on the observations in cluster $C_{\Omega\omega} \cap C_j^0$, $\omega = 1, ..., \Omega$, $j = 1, ..., \Omega_0$. Let

$$RSS_\omega = \left\| \widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{X}_{C_{\Omega\omega}} \widehat{\boldsymbol{\beta}}_{\Omega\omega} \right\|^2$$

denote the sum of the squares of the residuals for the $C_{\Omega\omega}$ cluster, and

$$RSS = RSS(\Omega) = \sum_{\omega=1}^{\Omega} RSS_\omega.$$

Define

$$F_N\left(\Pi_\omega^{(N)}\right) = N \log\left(\frac{RSS}{NT}\right) + f(\Omega)\,\theta_N, \tag{9}$$

where $f(\Omega)$ is a strictly increasing function of Ω and θ_N is a sequence of constants the size of which depends on N. For example, we often take $\theta_N = \sqrt{N}$ and f as the identity function. We propose estimating the number of clusters and the corresponding partition by minimising the following objective function:

$$F_N\left(\Pi_{\widehat{\Omega}_0}^{(N)}\right) = \min_{1 \le \Omega \le \xi} \min_{\Pi_\Omega^{(N)}} F_N\left(\Pi_\Omega^{(N)}\right), \tag{10}$$

where $\widehat{\Omega}_0$ is the value of Ω that minimises F_N. It will be shown in the following section that, under certain conditions, this criterion identifies Ω_0 with probability one as N grows large.

Using the above criterion to compare two distinct partitions we have

$$\begin{aligned} & F_N\left(\Pi_\Omega^{(N)}\right) - F_N\left(\Pi_{\Omega_0}^{(N)}\right) \\ &= N \log\left[1 + \frac{RSS(\Omega) - RSS(\Omega_0)}{RSS(\Omega_0)}\right] + \left[f(\Omega) - f(\Omega_0)\right]\theta_N, \\ &\sim \frac{RSS(\Omega) - RSS(\Omega_0)}{RSS(\Omega_0)/N} + \left[f(\Omega) - f(\Omega_0)\right]\theta_N. \end{aligned}$$

The residual sum of squares for the Π_{Ω_0} partition divided by N, which appears in the denominator of the ratio in the expression above, is a measure of the variability in the data. Thus, heuristically, the first term compares the goodness of fit of a model normed by a measure of the overall level of spread. Therefore, the proposed criterion is invariant to the scale of the data. This is important because in practice for any fixed N and T, multiplying the variables by a constant scalar will change the RSS value without altering the estimates of the slope parameters. The second term in the above expression is a penalty for overfitting, which reflects the fact that the minimum RSS of the estimated model is monotone decreasing in the number of

clusters and therefore it tends to over-parameterise the model by allowing for more clusters than may actually exist. Hence, the penalty acts essentially as a filter to ensure that the preferred clustering outcome partitions between rather than within clusters. The intuition of the procedure is identical to a standard model selection criterion, although the study of the asymptotics is more complicated in the present case because the number of individuals contained in a given cluster may vary with N.

3 Asymptotic Properties of Clustering Criterion

The following assumptions are required to establish the asymptotic properties of the proposed clustering criterion:

A.1 There exists a fixed constant, $0 < c_\omega < 1$, with $\sum_{\omega=1}^{\Omega_0} c_\omega = 1$, such that $N_\omega/N \to c_\omega$ for $\omega = 1, ..., \Omega_0$, as $N \to \infty$.

A.2 Ω_0 is a fixed unknown integer, such that $0 < \Omega_0 \leq \xi$, where ξ is fixed and known.

A.3 Given the covariates $X_{\omega it}$ corresponding to the observations in the ω^{th} cluster, the error vectors $\boldsymbol{\varepsilon}_{\omega i} = (\varepsilon_{\omega i1}, ..., \varepsilon_{\omega iT})'$ for the individuals in the cluster are independent and identically distributed random vectors with mean vector $\mathbf{0}$ and for some $\delta > 0$, $E\left|\varepsilon_{\omega it}\right|^{2+\delta} < \infty$. To avoid trivialities assume some elements of $\boldsymbol{\varepsilon}_{\omega i}$ have non-zero variance.

Let C_ℓ denote a true class or a subset of a true class with N_ℓ elements. Given the matrix $\widetilde{X}_{C_\ell}$, let $\widetilde{X}_{C_\ell}^{(t)}$ be the submatrix consisting of rows $t, t+T, ..., t+(N_\ell - 1)T$ of $\widetilde{X}_{C_\ell}$ for $t = 1, ..., T$.

A.4 There exist constants $\alpha_1 > 0$ and $\alpha_2 > 0$ such that the eigenvalues of $N_\ell^{-1}\widetilde{X}_{C_\ell}'\widetilde{X}_{C_\ell}$ and $N_\ell^{-1}\widetilde{X}_{C_\ell}^{(t)\prime}\widetilde{X}_{C_\ell}^{(t)}$ lie in $[\alpha_1, \alpha_2]$ for N_ℓ large enough.

A.5 For any column vector $\mathbf{x}_{\omega\ell}$ of $\widetilde{X}_{C_\ell}$, its elements $x_{\omega\ell}^{(1)}, ..., x_{\omega\ell}^{(N_\ell T)}$ satisfy the condition

$$\sum_{i=1}^{N_\ell T}\left|x_{\omega\ell}^{(i)}\right|^{2+\delta} = O_p\left[\left(\mathbf{x}_{\omega\ell}'\mathbf{x}_{\omega\ell}\right)^{(2+\delta)/2} / \log\left(\mathbf{x}_{\omega\ell}'\mathbf{x}_{\omega\ell}\right)^{1+\delta}\right] \tag{11}$$

for $1 \leq \omega \leq \Omega$ and some $\delta > 0$.

A.1 ensures that no clusters are asymptotically negligible. In particular, it implies that for the true partition there exist fixed constants $d_\omega \in (0,1)$ such that $d_\omega < \frac{N_{0\omega}}{N} < 1$, $\omega = 1, ..., \Omega_0$ for N large enough. Assumption A.2 ensures that the total number of clusters is bounded by a known integer, ξ.[3] Assumption A.3 is common in panel data models and implies that the covariates are strictly exogenous with respect to the idiosyncratic error component, $\boldsymbol{\varepsilon}_{\omega i}$, although not with respect to the total error term. Observe also that $\varepsilon_{\omega it}$ is permitted to be serially correlated in an arbitrary way and heteroskedastic across clusters and over t. Assumptions A.4−A.5 describe the behaviour of the covariates. A.4 is employed for identification purposes and ensures that $\left(N_\ell^{-1}\widetilde{X}_{C_\ell}'\widetilde{X}_{C_\ell}\right)^{-1}$ exists in probability for all N_ω sufficiently large.

For any set C_ℓ which is a true cluster, a subset of a true cluster or a union of subsets of a true cluster with $|C_\ell| = N_\ell$, let $P_{\widetilde{X}_{C_\ell}}$ denote the projection matrix

$$P_{\widetilde{X}_{C_\ell}} = \widetilde{X}_{C_\ell}\left(\widetilde{X}_{C_\ell}'\widetilde{X}_{C_\ell}\right)^{-1}\widetilde{X}_{C_\ell}', \tag{12}$$

based on the corresponding $\widetilde{X}_{C_\ell}$ matrix. Let $\boldsymbol{\varepsilon}_{C_\ell}$ denote the vector of corresponding error terms. The following lemma controls the rate of growth of a weighted sum of random variables.

Lemma 1 *Let $\varpi_1, \varpi_2, ...$ be a sequence of independent random variables with zero mean, such that $0 < E(\varpi_i^2) = \sigma_i^2$ and $E|\varpi_i|^{2+\delta} < \tau < \infty$ for some $\tau > 0$, $\delta > 0$ and $i = 1, 2, ...$ Furthermore, let $\alpha_1, \alpha_2, ..., \in \mathbb{R}$ be a sequence of constants such that*

$$\begin{aligned} (i) \quad B_N^2 &= \sum_{i=1}^N \alpha_i^2 \to \infty; \\ (ii) \quad \sum_{i=1}^N |\alpha_i|^{2+\delta} &= O_p\left\{B_N^{2+\delta}\left(\log B_N^2\right)^{-1-\delta}\right\}, \text{ for some } \delta > 0. \end{aligned}$$

Then, for $N \to \infty$

$$T_N = \sum_{i=1}^N \alpha_i \varpi_i = O\left(\left(B_N^2 \log\log\left(B_N^2\right)\right)^{\frac{1}{2}}\right) \ a.s.$$

Proof. See Shao and Wu (2005), Lemma 3.5. ■

Write

$$\boldsymbol{\varepsilon}_{C_\ell} = \boldsymbol{\nu}_{C_\ell}^{(1)} + ... + \boldsymbol{\nu}_{C_\ell}^{(T)}, \tag{13}$$

[3] The choice of ξ is empirically immaterial, as explained in Section 4.1.

where the i^{th} element of $\boldsymbol{\nu}_{C_\ell}^{(t)}$ is $(\boldsymbol{\varepsilon}_{C_\ell})_i I(i \in \{t, t+T, t+2T, ...\})$. For example, $\boldsymbol{\nu}_{C_\ell}^{(1)} = (\varepsilon_{\omega 11}, 0, ..., 0, \varepsilon_{\omega 21}, 0, ..., \varepsilon_{\omega N_\ell 1}, 0, ..., 0)'$ and so on. The non-zero elements of the vector $\boldsymbol{\nu}_{C_\ell}^{(t)}$ are the $i.i.d.$ error terms corresponding to the observations at time t for the elements in the cluster. We can write

$$\boldsymbol{\varepsilon}'_{C_\ell} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = \sum_{t=1}^{T} \sum_{s=1}^{T} \boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}. \tag{14}$$

Using the idempotent nature of the matrix $P_{\widetilde{X}_{C_\ell}}$ and the Cauchy-Schwartz inequality we have

$$\begin{aligned} \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)^2 &= \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}}^2 \boldsymbol{\nu}_{C_\ell}^{(s)}\right)^2 \\ &= \left(\left(P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)}\right)' \left(P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)\right)^2 \\ &\leq \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)}\right) \left(\boldsymbol{\nu}_{C_\ell}^{(s)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right). \end{aligned} \tag{15}$$

Thus, if $\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)} = O(\log\log N_{C_\ell})$ a.s. for each t, then $\boldsymbol{\varepsilon}'_{C_\ell} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = O(\log\log N_{C_\ell})$ a.s.

Applying Lemma 1 along with assumptions CA.1-CA.3 we have

$$\nu_{C_\ell}^{(t)\prime} \widetilde{X}_{C_\ell} = O\left(N_\ell \log\log N_\ell\right)^{\frac{1}{2}} \text{ a.s.} \tag{16}$$

Therefore,

$$\boldsymbol{\varepsilon}'_{C_\ell} \widetilde{X}_{C_\ell} = O\left(\left(N_\ell \log\log N_\ell\right)^{\frac{1}{2}}\right) \text{ a.s.} \tag{17}$$

Furthermore, A.4 ensures that the elements of $\left(\widetilde{X}'_{C_\ell} \widetilde{X}'_{C_\ell}\right)^{-1}$ are $O\left(N_\ell^{-1}\right)$. Hence, using (16) and arguing as in the proof of Lemma A.2 of Bai, Rao and Wu (1999) we have

$$\begin{aligned} \boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)} &= \boldsymbol{\nu}_{C_\ell}^{(t)\prime} \widetilde{X}_{C_\ell} \left(\widetilde{X}'_{C_\ell} \widetilde{X}_{C_\ell}\right)^{-1} \widetilde{X}'_{C_\ell} \boldsymbol{\nu}_{C_\ell}^{(t)} \\ &= O\left(\log\log N_\ell\right) \text{ a.s.} \end{aligned} \tag{18}$$

As a result,

$$\boldsymbol{\varepsilon}'_{C_\ell} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = O\left(\log\log N_\ell\right) \text{ a.s.} \tag{19}$$

The results in (17) and (19) are key to proving that the clustering algorithm converges to the true number of clusters. The asymptotics are developed by considering class growing sequences. That is, we will assume that as N increases, the sequence of true partitions of $\{1, 2, ..., N\}$ is naturally nested, i.e.

$$C_{\Omega\omega}^{(N)} \subseteq C_{\Omega\omega}^{(N+1)} \text{ for all } \omega = 1, ..., \Omega, \text{ for large } N. \tag{20}$$

In other words, the asymptotics can be conceived via a 'class-growing sequence' approach, which assigns the $(N+1)^{th}$ observation to any cluster of the previous partition based on the first N observations. The following theorem shows that the criterion in (10) selects the true number of clusters amongst all class-growing sequences with probability one for N large enough.

Theorem 2 *Let* $\lim_{N\to\infty} N^{-1}\theta_N = 0$ *and* $\lim_{N\to\infty} (\log\log N)^{-1}\,\theta_N = \infty$. *Suppose that assumptions A.1-A.5 hold and* Π_{Ω_0} *is the true clustering partition corresponding to model* (5). *Then the clustering criterion in* (10) *is strongly consistent – that is, it selects* Ω_0*, the true number of clusters among all class-growing sequences, with probability one as* $N \to \infty$.

Proof. See Appendix. ■

The first condition in Theorem 2 prevents estimating too many clusters asymptotically while the second condition prevents under-fitting. Similar conditions underlie well-known model selection criteria such as the AIC and the BIC, except that the criterion above is developed for the purpose of clustering individuals. Our class-growing approach is motivated by Shao and Wu (2005), who prove consistency of a similar criterion function for the cross-sectional regression model. Our model is more general, while it permits cross-sectional dependence in the errors and arbitrary forms of residual serial correlation. Moreover, the proposed criterion is invariant to the scale of the data.

In practice the unknown Φ in the transformation matrix Q_ω can be replaced by any consistent estimator $\widehat{\Phi}$ for fixed T. For example, given that the covariates are strictly exogenous with respect to the purely idiosyncratic error, $\boldsymbol{\varepsilon}_{\omega i}$, $\widehat{\Phi}$ can be obtained using the method of Pesaran (2006), or using principal components analysis based solely on the covariates. Sarafidis and Wansbeek (2011) provide an overview of these procedures.

4 Implementation

The number of ways to partition a set of N objects into Ω nonempty subsets is given by a 'Stirling number of the second kind', which is one of two types of Stirling numbers that commonly occur in the field of combinatorics.[4] Stirling numbers of

[4] See, for example, Rota (1964).

the second kind are given by the formula

$$S(N,\Omega)=\frac{1}{\Omega!}\sum_{\omega=0}^{\Omega}(-1)^{\Omega-\omega}\left(\begin{array}{c}\Omega\\ \omega\end{array}\right)\omega^{N}. \tag{21}$$

Therefore, the total number of partitions is exponential in N and, in fact, the optimization problem becomes intractable even for relatively small values of N and Ω. To see the order of the magnitude of a Stirling number, for $N=50$ and $\Omega=3$ the total number of distinct partitions is larger than $1.19\cdot10^{23}$. This implies that if we assumed, rather optimistically, that a given computer was able to estimate $10,000$ panel regressions every second, one would require about $3.79\cdot10^{11}$ years to exhaust all possible partitions. Clearly, a global search over all possible partitions is not feasible, even with small data sets. To deal with this issue, we propose a partitional algorithm based on K-means clustering.

4.1 K-means regression clustering

K-means algorithms are common in partitional cluster analysis (see, e.g., Everitt, 1993, and Kaufman and Rousseeuw, 1990). The algorithm we adopt in this paper is suitable for regression clustering and it can be outlined in the following steps[5]:

1. Given an initial partition and a fixed number of clusters, estimate the model for each cluster separately and calculate RSS;

2. Assign the i^{th} cross-section to all remaining clusters and obtain the resulting RSS value that arises in each case. Finally, assign the i^{th} individual into the cluster that achieves the smaller RSS value;

3. Repeat the same procedure for $i=1,...,N$;

4. Repeat steps 2-3 until RSS cannot be minimised any further.

Once the partition that achieves the minimum RSS value has been determined, one may repeat steps 1-4 for different numbers of clusters. The final number of clusters can be determined by the value that minimises

$$N\log\left(\frac{RSS}{NT}\right)+f(\Omega)\,\theta_{N}, \tag{22}$$

[5]The algorithm is written as an ado file in Stata 11 and it will be made available to all Stata users on the web.

where $f(\Omega)$ is a strictly increasing function of Ω and θ_N is chosen such that it satisfies the bounds in Theorem 2.[6]

A simple initial choice is to set $f(\Omega) = \Omega$ and $\theta_N = \sqrt{N}$, which lies between the lower and upper bounds set out in Theorem 2. These values have been found to be reliable across a range of models in simulations. Further parametrisations for the penalty function are discussed in the next section.

The basic idea of steps 2-4 of the algorithm is very similar to that underlying steepest descent algorithms used to solve non-linear optimization problems. In particular, this type of algorithm starts at an initial point and then generates a sequence of moves from one point to another, each leading to an improved value of the objective function, until a local minimum is reached. The local minimum is the partition that minimises the within-cluster residual sum of squares, $\sum_{\omega=1}^{\Omega} \left\| \widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{X}_{C_{\Omega\omega}} \widehat{\boldsymbol{\beta}}_{\Omega\omega} \right\|^2$. Using the properties of least squares residuals one can write

$$\sum_{\omega=1}^{\Omega} \left\| \widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{\mathbf{X}}_{C_{\Omega\omega}} \widehat{\boldsymbol{\beta}}_{p} \right\|^2 = \sum_{\omega=1}^{\Omega} \left\| \widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{\mathbf{X}}_{C_{\Omega\omega}} \widehat{\boldsymbol{\beta}}_{\Omega\omega} \right\|^2 + \sum_{\omega=1}^{\Omega} \left\| \widetilde{\mathbf{X}}_{C_{\Omega\omega}} \left(\widehat{\boldsymbol{\beta}}_{\Omega\omega} - \widehat{\boldsymbol{\beta}}_{p} \right) \right\|^2,$$

where $\widehat{\boldsymbol{\beta}}_p$ denotes the pooled estimator. Since the term on the left-hand side remains constant across all possible partitions it is easy to see that minimising the within-cluster residual sum of squares is equivalent to maximising the between cluster squared differences of the fitted values obtained from the cluster-specific estimates versus the pooled estimate.

Notice that in each move the assignment of N individuals into Ω clusters entails $N\Omega$ regressions and $N(\Omega - 1)$ comparisons of residual sums of squares. The convergence of the algorithm to a local minimum is guaranteed (see Selim and Ismail, 1984, for a proof). Intuitively, this is because the method alters a given partition only if assigning an individual to a different cluster leads to a lower residual sum of squares. Therefore, the algorithm cannot choose a partition that was abandoned at an earlier stage. Thus, since each partition is generated at most once and the number of partitions is finite, the algorithm is finitely convergent. The time complexity of the algorithm is proportional to $KN\Omega B$, where B is the number of iterations and the value of which depends on the distribution of the data points. The simulation experiments we have performed indicate that five

[6]Notice that the choice of ξ in the algorithm is immaterial because if the chosen value of ξ is smaller than Ω_0, the number of the clusters minimising the criterion function will equal the maximum number of clusters allowed, which will indicate that the researcher should allow for more clusters.

iterations, or less, typically suffice and only rarely more than ten iterations are required. Of course, convergence to the global minimum requires, in addition, that steps 2-4 are reiterated using a sufficiently large number of random starts to escape local minima. Alternatively, the initial partition can be chosen carefully based on a set of observed attributes, such as the individual-specific estimated slope coefficients, or a set of variables that do not enter directly into the model. This possibility is studied in the following section.

4.2 Choosing the Initial Partition

There are several ways to choose the initial partition. For the case where there is a single variable to which slope parameter heterogeneity can be attributed, one can use the property that when the cross-sectional units are ordered according to the value of this variable, the partition that minimises the objective function, total RSS, is a contiguous partition, i.e. each cluster corresponds to a single interval that is disjoint from all other clusters; see Fisher (1958). This reduces the number of possible partitions from a Stirling number of the second kind to the binomial coefficient $\begin{pmatrix} N-1 \\ \Omega-1 \end{pmatrix}$. This result comes from the fact that there are $N-1$ intervals defined by the N ordered elements of the individual-specific slope coefficients, which are segmented by $\Omega-1$ dividors. The number of ways of choosing $\Omega-1$ division points on $N-1$ intervals yields the total number of possible contiguous partitions. Thus, the computational complexity of solving the optimization problem is $O\left(N^{\Omega}\right)$ and so for fixed Ω it is polynomial. Using the same example as before, for $N=50$ and $\Omega=3$ the total number of distinct contiguous partitions equals $1,176$, which implies a reduction of 10 orders of magnitude.

Unfortunately, the above procedure becomes unappealing for N moderately large and $\Omega>3$. One would benefit from a more efficient solution algorithm which exploits the additive property of residual sum of squares and is polynomial in N and independent of Ω. Hence we develop an iterative algorithm based on a dynamic programming formulation, which solves the problem into polynomial time, or more specifically in $O\left(N^2\Omega\right)$. The objective is to partition the contiguous set into at most Ω non-overlapping clusters so as to minimise RSS. Before giving the formal algorithm we calculate an $N\times N$ matrix of the RSS function defined as follows: $RSS\left(i,j\right)$ is the residual sum of squares for individuals i, $i+1, i+2, ..., j$, for $1\leq i\leq j\leq N$. Clearly, computing all values in the matrix requires $O\left(N^2\right)$ time. Since the number of clusters is bounded from above, a forward dynamic

programming algorithm proves more appropriate for our optimization problem.

We define the following two-state RSS function $f(r,m)$ where r describes the last individual that has been assigned to a cluster and m is the number of clusters used for the first $1,2,...,r$ individuals. Thus, we assume that individuals $r+1,r+2,...,N$ have been optimally assigned into clusters. Our decision variable is given by r' and describes the last individual not included in the current cluster. Hence we choose to include in the current cluster $r'+1,r'+2,...,r$ individuals. We begin with our objective

$$\min\left\{f(N+1,m)\right\}.$$

The boundary condition is

$$f(0,0)=0.$$

Next we define the following recursive relation between $f(r,m)$ and $f(r',m')$, whereas m' is the number of clusters for individuals $1,...,r'$:

$$f(r,m)=\min_{\substack{1\leq r'<r\\ m\leq r}}\left\{f(r',m-1)+RSS(r'+1,...,r)+g(m)\right\},$$

whereas

$$g(m)=\begin{cases}\infty \text{ if } m>\Omega\\ 0 \text{ elsewhere.}\end{cases}$$

Notice that m' is always equal to $m-1$ by definition. In addition, we require $g(m)$ to ensure that we do not create more than Ω clusters. The first term in the recursive equation is the minimum RSS value when assigning the first r individuals into the $m-1$ cluster, while the second term is the RSS value of a single cluster containing individuals $r'+1,...,r$. The algorithm stores the minimum residual sum of squares for each m (at most Ω values), which can then be used as inputs for the model selection criterion in (22).

The running time of the algorithm is $O(N^2\Omega)$. For a fixed Ω this is clearly of order $O(N^2)$, where as for a general Ω it is $O(N^3)$ since Ω is bounded by N. The number of possible states is $O(N\Omega)$ since we consider N individuals and Ω clusters. At each state there are at most N alternatives since the state variable m grows by 1 where as the r variable has at most $O(N)$ alternatives. The exact number of calculations for the $RSS(i,j)$ matrix is $N(N+1)/2$. The recursive function is computed exactly $N+\sum_{\omega=1}^{\Omega}(N-\omega+1)(N-\omega)/2$ times.

Unfortunately the problem of obtaining the initial partition becomes much more complicated when there exists more than one variable to which slope parameter heterogeneity can be attributed. Suppose, for example, that one wishes to

obtain the initial partition based on optimal clustering of the individual-specific estimated slope coefficients; in the multivariate case this is not straightforward because the ordering of the cross-sectional units can vary across different variables. One possibility is to convert the problem into a set partitioning problem and then build up an algorithm for multi-dimensional data clustering (see e.g. Beasley and Chu, 1996, and Wan, Wong and Prusinkiewicz, 1988). In particular, the problem can be stated in the following way. There are N individuals and each of them contains a vector, the entries of which are different variables. Following Rao (1971) we use the following 'string condition': "in an optimal solution, each group should consist of at least one individual unit, which for convenience will be denoted as the leader of the group, such that the distance between the leader and any individual that does not belong to the same group is not less than the distance between the leader and any individual within the same group." Mathematically this can be expressed as

$$d_{i,j\in g_i} \leq d_{i,j\notin g_i},$$

where $d_{i,j}$ is the Euclidian distance between individuals i and j, and g_i is the group the leader of which is individual i. Notice that this condition is different from the property of contiguity, adopted in the dynamic programming algorithm analysed above, in that clusters do not necessarily consist of consecutive points on the real line measuring a single observed variable. Therefore, since there are N individuals and each of these is a candidate to be a leader of a group, the string condition implies the existence of $N(N-1)+1$ groups, including the one comprising all individuals. This can be seen if we let $j_1, j_2, ..., j_{n-1}$ be entities such that

$$d_{i,i} = 0 < d_{j_1,i} < d_{j_2,i} < ..., d_{j_{n-1},i}.$$

Thus, the problem takes the form

$$\min CY$$

$$\begin{aligned} \text{subject to } AY &= b \\ Y_i &\in \{0,1\}, \end{aligned}$$

where C is a $N(N-1)+2$ row vector that contains the cost for a particular grouping, which in our context is the cluster-specific RSS, Y is a $N(N-1)+2$ column vector representing whether a particular grouping is utilised or not in the optimal solution, A is a $(N+1)\times[N(N-1)+2]$ matrix and b is a $N+1$

column vector given by $(1, ..., \Omega)$. Each column of A, except the last one, reflects a possible grouping of the N individuals and each row corresponds to an individual. The last column of A has all zeros except one in the last row which restricts the total number of clusters to be at most Ω. Notice that some of the groupings may be identical and therefore should be deleted. Further reductions are suggested by Garfinkel and Nemhauser (1969) and Beasley and Chu (1996). The final procedure provides the optimal Ω non-overlaping clusters subject to the string condition mentioned above. Both our dynamic programming algorithm and our set-partitioning algorithm will be made available on the web.

5 Simulation Study

In this section we carry out simulation experiments to investigate the performance of our criterion in finite samples. Our main focus lies on the choice of θ_N and the effect of (i) the number of clusters, (ii) the size of N, (iii) the number of regressors and (iv) the signal-to-noise ratio in the model. We also pay attention to the properties of the estimators that arise from the estimated partitions, as well as the pooled OLS and FE estimators.

5.1 Experimental Design

The underlying process is given by

$$\begin{aligned} y_{\omega it} &= \sum_{k=1}^{K} \beta_{k\omega} x_{k\omega it} + \eta_{\omega i} + u_{\omega it}, \\ t &= 1, ..., T,\ i\,[\in \omega] = 1, ..., N_\omega \text{ and } \omega = 1, ..., \Omega_0, \end{aligned} \tag{23}$$

where $\eta_{\omega i}$ is drawn in each replication from $i.i.d.N\left(0, \sigma_\eta^2\right)$, while $x_{k\omega it}$ is drawn from $i.i.d.N\left(\mu_{x_{k\omega}}, \sigma_{x_{k\omega}}^2\right)$. $u_{\omega it}$ obeys a single-factor structure

$$u_{\omega it} = \lambda_{\omega i}\phi_t + \varepsilon_{\omega it}, \tag{24}$$

where $\lambda_{\omega i} \sim i.i.d.N\left(0, 0.5\sigma_u^2\right)$, $\phi_t \sim i.i.d.N\left(0, 1\right)$ and $\varepsilon_{\omega it} \sim i.i.d.N\left(0, 0.5\sigma_u^2\right)$, such that $Var\left(u_{\omega it}\right) = \sigma_u^2$.

Define $y_{\omega it}^* = y_{\omega it} - \eta_{\omega i}$, such that (23) can be rewritten as

$$y_{\omega it}^* = \sum_{k=1}^{K} \beta_{k\omega} x_{k\omega it} + u_{\omega it}, \tag{25}$$

and let the signal-to-noise ratio be denoted by $\zeta_\omega = \sigma^2_{s_\omega}/\sigma^2_{u_\omega}$, where $\sigma^2_{s_\omega}$ and $\sigma^2_{u_\omega}$ denote the variance of the signal and noise, respectively, for the ω^{th} cluster. $\sigma^2_{s_\omega}$ equals

$$\sigma^2_{s_\omega} = var\left(y^*_{\omega it} - u_{\omega it}\right) = var\left(\sum_{k=1}^{K} \beta_{k\omega} x_{k\omega it}\right) = \sum_{k=1}^{K} \beta^2_{k\omega} \sigma^2_{x_{k\omega}}. \tag{26}$$

This implies that for a given value of $\left\{\sigma^2_{x_{k\omega}}\right\}_{k=1}^{K}$ and $\sigma^2_{u_\omega}$, the signal-to-noise ratio for the ω^{th} cluster depends on the value of $\{\beta_{k\omega}\}_{k=1}^{K}$. Thus, for example, scaling the coefficients upwards by a constant factor will increase ζ and this may improve the performance of the model selection criterion; however, there is no natural way to choose the value of such scalar. Furthermore, notice that for fixed $\sigma^2_{u_\omega}$ alternating K will change $\sigma^2_{s_\omega}$ and thereby the performance of the criterion may also be affected. We control both these effects by normalising $\sigma^2_{u_\omega} = 1$, $\zeta_\omega = \zeta$, for $\omega = 1, ..., \Omega_0$ and setting $\sigma^2_{x_{k\omega}} = \zeta/\left(\beta^2_{\omega k} K\right)$. In this way, the signal-to-noise ratio in our design is invariant to the choice of K and the scale of $\{\beta_{k\omega}\}_{k=1}^{K}$. The values of the slope coefficients are listed in Table 1. We consider $\zeta = \{4, 8\}$, $N = \{100, 400\}$ with $T = 10$, $K = \{1, 4\}$ and $\Omega_0 = \{1, 2, 3\}$.[7] We set $N_1 = 0.7N$, $N_2 = 0.3N$ for $\Omega_0 = 2$ and $N_1 = 0.4N$, $N_2 = 0.3N$, $N_2 = 0.3N$ for $\Omega_0 = 3$. This allows the size of the clusters to be different. We perform 500 replications in each experiment. To reduce the computational burden, we fit models with $\Omega = 1, 2, 3$ clusters when $\Omega_0 = 1$, $\Omega = 1, 2, 3, 4$ clusters when $\Omega_0 = 2$ and $\Omega = 1, 2, 3, 4, 5$ clusters when $\Omega_0 = 3$.

ϕ_t is estimated in each replication based on the method of Pesaran (2006) and the model is orthogonalised prior to estimation by premultiplying the $T \times 1$ vectors of observed variables, $\mathbf{y}_{\omega i} = (y_{\omega i1}, ..., y_{\omega iT})'$ and $\mathbf{x}_{k\omega i} = (x_{k\omega i1}, ..., x_{k\omega iT})'$ for $k = 1, ..., K$, by the $T \times T$ idempotent matrix $M = I_T - \overline{Z}\left(\overline{Z}'\overline{Z}\right)^{-1}\overline{Z}'$, $\overline{Z} = \left(\overline{Z}_1, ..., \overline{Z}_T\right)'$, with typical entry $\overline{Z}_t = N^{-1}\sum_{i=1}^{N} \mathbf{z}_{it}$, $\mathbf{z}_{it} = (y_{it}, \mathbf{x}'_{it})'$.

[7] We also set $\mu_{x_{k\omega}} = 1$ for $k = 1, ..., K$ and $\omega = 1, ..., \Omega$.

Table 1. Parameter values used in the simulation study.

	$K=1$	$K=4$
$\Omega_0 = 1$	$\beta = 1$	$\boldsymbol{\beta} = \begin{pmatrix} 1 \\ .5 \\ .75 \\ 2 \end{pmatrix}$
$\Omega_0 = 2$	$\beta_1 = 1$ $\beta_2 = .5$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 1 \\ .5 \\ .75 \\ 2 \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} .5 \\ .25 \\ .375 \\ 1 \end{pmatrix}$
$\Omega_0 = 3$	$\beta_1 = 1$ $\beta_2 = .5$ $\beta_3 = -.25$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 1 \\ .5 \\ .75 \\ 2 \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} .5 \\ .25 \\ .375 \\ 1 \end{pmatrix}, \boldsymbol{\beta}_3 = \begin{pmatrix} -.25 \\ 1 \\ 1.5 \\ 0.5 \end{pmatrix}$

5.2 Results

Tables A1-A3 in the appendix report the results of our simulation experiments in terms of the relative frequency of selecting Ω clusters when the true number of clusters is Ω_0. The relative frequency of selecting the true number of clusters is emphasised in bold. Since the property of consistency of $\widehat{\Omega}$ only requires that $f(\Omega)$ is stricly increasing in Ω and θ_N satisfies $\lim_{N\to\infty} N^{-1}\theta_N = 0$ and $\lim_{N\to\infty} (\log\log N)^{-1}\theta_N = \infty$, there is a broad range of values for the penalty function one can choose from. In this study we set $f(\Omega) = \Omega$ such that

$$MIC_j = N \log\left(\frac{RSS}{NT}\right) + \Omega\theta_j \text{ for } j = 1, ..., 4,$$

where $\theta_1 = 2$, $\theta_2 = \log N$, $\theta_3 = (1/\psi)\left[(\log N)^{\psi} - 1\right]$ and $\theta_4 = \sqrt{N}$. MIC_1 and MIC_2 resemble the Akaike and Bayesian information criteria, respectively, except that they are applied to the clustering selection problem. θ_3 is motivated from the fact that $(1/\psi)\left[(\log N)^{\psi} - 1\right] \to \log\log N$ as $\psi \to 0$ and hence for any ψ bounded away from zero, the lower bound of Theorem 2 is satisfied. As a rule of thumb, we choose ψ such that θ_3 lies between θ_2 and θ_4, in particular, we set $\theta_3 = w\theta_2 + (1-w)\theta_4$, $w = 1/3$. An alternative method of selecting the value of ψ can be based on the following heuristic algorithm. Firstly, for a given number of clusters a parametric boostrap algorithm is run B times – that is, the responses are sampled using the optimal partition obtained for this particular cluster number.[8] Subsequently, an interval of ψ is determined such that the correct number of clusters is selected in all bootstrapped samples. After repeating this procedure

[8] We set $B = 50$.

Ω times, the intersection of all intervals is chosen as our admissible ψ values. It is worth noting that in the present study the value of ψ obtained from the rule of thumb almost always lies in the admissible set of values of ψ. Therefore, in what follows we report results for this particular θ_3 value only. It is also possible to determine ψ using cross-validation, by interpreting this as a smoothing parameter in a nonparametric regression. For each particular value of $\psi = \psi_g$, evaluated at a given g, where $g = 1, ..., G$, the optimal predicted value for $\mathbf{y}_i$, $\widehat{\mathbf{y}}_{-i}\left(\psi_g\right)$, is computed by dropping individual i from the sample, estimating the optimal number of clusters as well as the corresponding partition, and allocating individual i to the cluster that achieves the lowest prediction error for $\mathbf{y}_i$. The optimal value of ψ is then determined by minimising the objective function, $S\left(\psi_g\right) = \sum_{i=1}^{n} \left|\left|\mathbf{y}_i - \widehat{\mathbf{y}}_{-i}\left(\psi_g\right)\right|\right|^2$, over a grid of values of ψ. This procedure, however, can be prohibitively time-consuming for moderately large N and as such we have not pursued this any further.

As we can see from the tabulated results, both MIC_3 and MIC_4 perform very well in all circumstances. This holds true for all values of N, K and ζ. Naturally, the performance of both criteria improves with larger values of N and ζ.[9] On the other hand, MIC_1 performs poorly in most circumstances in that it constantly overestimates the true number of clusters. This is not surprising as the criterion is not consistent for N large. In fact, its performance deteriorates as N increases. MIC_2 is a special case of our criterion and performs somewhat better than MIC_1. Notwithstanding, in a lot of cases it largely overestimates the number of clusters, especially when $\Omega_0 = 1, 2$. We have explored further the underlying reason for this result. We found that a larger penalty is required in the clustering regression problem to prevent over-fitting than what is typically used in the standard model selection problem.

Table A4 in the appendix reports the average point estimates of the parameters for $K = 1$.[10] Standard deviations are reported in parentheses. $\widehat{\boldsymbol{\beta}}_p$'denotes the pooled estimator that arises by pooling all clusters together, i.e. ignoring cluster-specific heterogeneity in the slope parameters. $\widehat{\boldsymbol{\beta}}_\omega$ denotes the estimator of the parameter vector for the ω^{th} cluster that arises from the estimated partition when Ω_0 is estimated using MIC_3. For $\widehat{\boldsymbol{\beta}}_p$ the true coefficient is taken to be the weighted average value of the cluster-specific unknown slope coefficients, with the weights

[9] ζ does not affect the results when $\Omega_0 = 1$ of course.

[10] To save space, we do not report the results obtained for $K = 4$ because similar conclusions can be drawn.

determined by the size of the true clusters. It is apparent that the bias in $\widehat{\boldsymbol{\beta}}_p$ is rather large. Its negative sign is due to the fact that the clusters with smaller coefficients exhibit relatively larger leverage because the variance of the regressors is larger for these clusters. In contrast, the cluster-specific estimators are virtually unbiased even if they are obtained from estimated clusters and the corresponding estimated partitions. This holds true even for $N = 100$, although the performance of the estimators naturally improves as N increases. In conclusion, we see that the criterion performs well, not only with respect to the estimate of Ω_0, but also in terms of leading to accurate cluster-specific coefficients.

6 Empirical Application

We apply the proposed partially heterogeneous framework on a cost function based on a panel data set of commercial banks operating in the United States. The issue of how to estimate scale economies and efficiency in the banking industry has attracted considerable attention among researchers due to the significant role that financial institutions play in economic prosperity and growth and, as a result, the major implications that these estimates entail for policy making.

6.1 Existing Evidence

In an earlier survey conducted by Berger and Humphrey (1997), the authors report more than 130 studies focusing on the measurement of economies of scale and the efficiency of financial institutions in 21 countries. They conclude that while there is lack of agreement among researchers regarding the preferred model with which to estimate efficiency and returns to scale, there seems to be a consensus on the fact that the underlying technology is likely to differ among banks. To this end, McAllister and McManus (1993) argue that the estimates of the returns to scale in the banking industry may be largely biased if one applies a single cost function to the whole sample of banks. This result is likely to remain even if one uses a more flexible functional form in the data, such as the translog form, because this would restrict, for example, banks of different size to share the same symmetric average cost curve. Hence, other interesting possibilities would be precluded, such as flat segments in the average cost curve over some ranges, or even different average cost curves among banks, depending on their size. Thus, the authors conclude:

> "These results, taken together, suggest that estimated cost functions

vary substantially depending on the range of bank sizes included in the sample. This extreme dependence of the results on the choice of the sample suggests that there are difficulties with the statistical techniques employed", page 389.

Similarly, Kumbhakar and Tsionas (2008) argue that since the banking industry contains banks of vastly different size, the underlying technology is very likely to be different across banks:

"The distribution of assets across banks is highly skewed. As a result of this, it is very likely that the parameters of the underlying technology (cost function in this case) will differ among banks", page 591.

Since this view appears to have been widely adopted in the banking literature, we estimate a partially heterogeneous cost regression model. A conceptually similar approach has been followed indirectly by Kaparakis et al (1994) and more recently by Kwan (2006), who distinguish between small and large banks and partition the population into two equally-sized sub-samples based on the median value of total assets. However, this partioning is rather arbitrary and there is no formal justification for imposing two clusters.

6.2 Methodology

The data set consists of a random sample of 551 banks, each observed over a period of 15 years. These data have been collected from the electronic database maintained by the Federal Deposit Insurance Corporation (FDIC).[11]

In the theory of banking there is not a univocal approach regarding one's view of what banks produce and what purposes they serve. In this paper we follow the "intermediation" approach, in which the banks are viewed as intermediators of financial and physical resources and produce loans and investments; see also Sealey and Lindley (1977). Under this approach, outputs are measured in money values and cost figures include interest expenses. The selection of inputs and outputs follows closely the study conducted by Hancock (1986). The variables used in the analysis are: c; the sum of the cost related to the three input prices that appear below, y_1; the sum of industrial, commercial and individual loans, real estate loans and other loans and leases, y_2; all other assets, p_l; the price of labour,

[11] See http://www.fdic.gov

measured as total expenses on salaries and employee benefits, divided by the total number of employees, p_k; the price of capital, measured as expenses on premises and equipment, divided by the dollar value of premises and equipment, and p_f; the price of loanable funds, measured as total expenses on interest, divided by the dollar value of deposits, federal funds purchased and other borrowed funds.

Hence, the model is specified as follows[12]:

$$\begin{aligned} c_{\omega it} &= \delta_{1\omega} y_{1,\omega it} + \delta_{2\omega} y_{2,\omega it} + \delta_{3\omega} p_{l,\omega it} + \delta_{4\omega} p_{k,\omega it} + \delta_{5\omega} p_{f,\omega it} + \upsilon_{\omega it}, \\ \upsilon_{\omega it} &= \alpha_{\omega i} + u_{\omega it},\ u_{\omega it} = \sum_{m=1}^{r} \lambda_{\omega i}^{m} \phi_t^m + \varepsilon_{\omega it}. \end{aligned} \tag{27}$$

The assumption of strict exogeneity of the regressors with respect to $\varepsilon_{\omega it}$ is standard in this context; see Kwan (2006), Kumbhakar and Tsionas (2008) and Fries and Taci (2005), among others. However, we deviate from the literature by allowing for cross-sectional dependence in the residuals, $u_{\omega it}$, by means of a multi-factor structure. These factors may capture distinct components of time-varying cost efficiency, or common shocks that hit the population of banks at time t. Since these unobserved common components are likely to be correlated with the regressors, strict/weak exogeneity with respect to $u_{\omega it}$ is violated, leading to biased and inconsistent parameter estimates.[13] We test for error cross-sectional dependence after estimating (27) allowing for a two-way error components model based on the fixed effects estimator. We use the test statistics developed by Pesaran (2004) and Pesaran, Ullah and Yamagata (2008) for this purpose. Both tests soundly reject the null hypothesis of no error cross-sectional dependence at the 5% level of significance. In particular, Pesaran's CD statistic equals 26.3 (p-value = 0.000) and the bias-adjusted LM statistic equals 116.9 (p-value = 0.000). Subsequently, we find two factors in the residuals based on the eigenvalue ratio test of Ahn and Horenstein (2008), and accordingly we orthogonalise all variables prior to estimation using principal components analysis.

6.3 Main Results

We cluster the sample of banks into up to six clusters based on our partitional clustering algorithm. The initial partition is chosen on the basis of bank size using the dynamic programming algorithm analysed in Section 4.2. This algorithm

[12] All variables are expressed in logs.

[13] A recent literature review on residual factor models is provided by Sarafidis and Wansbeek (2010).

finds the global minimum for a given number of clusters, i.e. the partition that minimises the within-cluster sum of squares of the deviations between each cross-sectional unit and the centroid of the cluster in which a particular cross-sectional unit belongs. Bank size is proxied by the fifteen-year average value of total assets for each individual bank.

Table 2 reports the values of MIC_j, $j = 1, ..., 4$, for $\Omega = 1, ..., 6$. As we can see, MIC_3 and MIC_2 suggest the presence of 5 clusters, while MIC_4, MIC_1, suggest four and six clusters, respectively. These finding corroborate the results of the simulation study, which show that under cross-sectional dependence MIC_4 might occasionally underestimate the true number of clusters while the penalty attached by MIC_1 is clearly insufficient to prevent over-fitting.

Table 2. Results for estimating the number of clusters.

Ω	1	2	3	4	5	6
MIC_1	-1086.2	-1227.0	-1249.0	-1260.2	-1270.4	**-1271.5**
MIC_2	-1082.5	-1219.5	-1237.9	-1245.4	**-1251.9**	-1249.3
MIC_3	-1079.6	-1213.8	-1299.2	-1233.9	**-1237.4**	-1231.9
MIC_4	-1072.7	-1200.0	-1208.6	**-1209.4**	-1203.0	-1190.7

Table 3 reports the estimation results obtained for model (27) for $\Omega = 5$. We adopt a notation similar to the simulation study; in particular, $\widehat{\boldsymbol{\beta}}_p$denotes the pooled estimator for the whole sample, $\widehat{\boldsymbol{\beta}}_\omega$ refers to the fixed effects estimate for the ω^{th} cluster and $\overline{\widehat{\boldsymbol{\beta}}}$ is the weighted average estimate of all clusters with the weights determined by the size of each estimated cluster. The clusters are sorted in ascending order such that cluster 1 contains on average the smallest banks and cluster 5 the largest banks.

Note that since our clustering procedure minimises the within-cluster residual sum of squares, the properties of the estimated standard errors obtained in the usual way are no longer known. Therefore, we use bootstrapping to attach a standard error to the estimated parameters. In particular, for each cluster we draw a "bootstrap sample" by sampling N times with replacement from the sample. We then estimate the parameters from the bootstrap sample, and we repeat this process $B\,(= 200)$ times, which provides estimates of the distribution one would get if one were able to draw repeated samples of N points from the unknown true distribution.

The results in Table 3 show that there are some large and statistically significant differences in the value of the regression coefficients across clusters. For

example, the estimated coefficient of the price of labour, $\widehat{\delta}_3$, appears to form a U-shaped function of size, which indicates that in a passage from small to medium sized banks economies of scale increase, and then decrease for large banks. In contrast, the estimated coefficient of loans, $\widehat{\delta}_1$, appears to rise as bank size increases, although it remains well below one. This implies that while there are increasing output returns for both small and large banks, the benefit of small banks getting larger is higher than for banks which are already large. It is worth mentioning that one draws qualitatively similar conclusions for $\Omega = 3, 4, 6$, and so the shape of these profiles appears to be robust to the choice of clusters. In summary, we see that banks of different size have different cost drivers and therefore pooling the data and imposing homogeneity in the slope parameters across the whole sample may yield misleading results. This becomes apparent when we compare $\widehat{\boldsymbol{\beta}}_p$ with $\overline{\widehat{\boldsymbol{\beta}}}$, the difference of which is statistically significant for most coefficients.

Table 3. Estimation Results[1,2]

	$\widehat{\delta}_1$	$\widehat{\delta}_2$	$\widehat{\delta}_3$	$\widehat{\delta}_4$	$\widehat{\delta}_5$
$\widehat{\boldsymbol{\beta}}_p$	.187	.221	.251	.007*	.548
	(.003)	(.005)	(.011)	(.004)	(.005)
$\widehat{\boldsymbol{\beta}}_1$	.021	.061	.309	.022	.401
	(.004)	(.009)	(.025)	(.007)	(.008)
$\widehat{\boldsymbol{\beta}}_2$	.049	.040	.226	.007*	.640
	(.006)	(.010)	(.021)	(.005)	(.008)
$\widehat{\boldsymbol{\beta}}_3$	.222	.096	.187	.006*	.497
	(.009)	(.014)	(.030)	(.013)	(.009)
$\widehat{\boldsymbol{\beta}}_4$	.117	.480	.005*	-.128*	.616
	(.015)	(.023)	(.047)	(.215)	(.021)
$\widehat{\boldsymbol{\beta}}_5$	.358	.236	.356	.056	.571
	(.011)	(.018)	(.047)	(.023)	(.021)
$\overline{\widehat{\boldsymbol{\beta}}}$	.110	.106	.243	.008	.530

1. Bootstrapped standard errors in parentheses.
2. '*' denotes non-significance at the 5% level.

7 Concluding Remarks

Slope parameter homogeneity versus heterogeneity is a topic that has intrigued research in the analysis of panel data since its infancy. In many cases the issue remains practically unresolved. In particular, a substantial body of empirical work

appears to suggest that while slope parameter homogeneity is usually rejected, the alternative of allowing these parameters to be individual-specific often leads to estimates with large standard errors, counterintuitive sign and inferior forecasting performance. This paper has proposed an intermediate modelling framework that imposes partially heterogeneous restrictions in the slope parameters. The unknown number of clusters and the corresponding partition are determined based on the concept of 'partitional clustering', using an information-based criterion that is strongly consistent for fixed T.

References

[1] Ahn, S. C., Horenstein, A. (2008). Eigenvalue ratio test for the number of factors. Mimeo.

[2] Bai, Z., Rao, C. R., Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference* 77:103-117.

[3] Baltagi, B. H., Griffin, J. M. (1997). Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline. *Journal of Econometrics* 77:303-327.

[4] Baltagi, B. H, Griffin, J. M, Xiong, W. (2000). To pool or not to pool: homogeneous versus heterogeneous estimators applied to cigaretter demand. *Review of Economics and Statistics* 82:117-126.

[5] Beasley, J. E, Chu P. C. (1996). A Genetic algorithm for the set covering problem. *European Journal of Operational Research* 94:392-404.

[6] Berger, A. N., Humphrey, D. B. (1997). Efficiency of financial institutions: international survey and directions for future research. *European Journal of Operational Research* 98:175-212.

[7] Browning. M., Carro, J. (2007). Heterogeneity and microeconometric modelling. In: Blundell, R., Newey, W., Persson, T. ed. *Advances in Economics and Econometrics* **3**. Cambridge: Cambridge University Press.

[8] Burnside, C. (1996). Production function regressions, returns to scale, and externalities. *Journal of Monetary Economics* 37:177-201.

[9] Durlauf, S., Johnson, P. (1995). Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics* 10:365–384.

[10] Everitt, B. (1993). *Cluster Analysis*. 3rd ed., London: Eward Arnold.

[11] Fernandez-Val, I. (2005). Bias correction in panel data models with individual specific parameters. Mimeo.

[12] Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association* 53: 789-798.

[13] Fries, S., Taci, A. (2005). Cost efficiency of banks in transition: evidence from 289 banks in 15 post-communist countries. *Journal of Banking and Finance* 29:55-81.

[14] Garfinkel, R. S., Nemhauser, G. L. (1969). The set-partitioning problem: set covering with equality constraints. *Operations Research* 17:848-856.

[15] Graham, B. S., Powel, J. L. (2008). Identification and estimation of 'irregular' correlated random coefficient models. Mimeo.

[16] Hancock, D. (1986). A model of financial firm with imperfect asset and deposit elasticities. *Journal of Banking and Finance* 10:37-54.

[17] Kapetanios, G. (2006). Cluster analysis of panel datasets using non-standard optimisation of information criteria. *Journal of Economic Dynamics and Control* 30:1389-1408.

[18] Kaparakis, E. I., Miller, S. M., Noulas, A. G. (1994). Short-run cost inefficiency of commercial banks: a flexible frontier approach. *Journal of Money, Credit and Banking* 26:875-893.

[19] Kaufman, L., Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis*. New York: John Wiley & Sons.

[20] Kumbhakar, S. C., Tsionas, E.G. (2008). Scale and efficiency measurement using a semiparametric stochastic frontier model: evidence from the U.S. commercial canks. *Empirical Economics* 34:585-602.

[21] Kwan, S. H. (2006). The X-efficiency of commercial banks in Hong Kong. *Journal of Banking and Finance* 30:1127-1147.

[22] McAllister, P. H., McManus, D. A. (1993). Resolving the scale efficiency puzzle in banking. *Journal of Banking and Finance* 17:389-405.

[23] Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels. University of Cambridge, Faculty of Economics, Cambridge Working Papers in Economics No. 0435.

[24] Pesaran, H. M. (2006). Estimation and inference in large heterogeneous panels With a multifactor error structure. *Econometrica* 74:967-1012.

[25] Pesaran, H. M., Shin, Y., Smith, R. J. (1999). Pooled mean group estimation of dynamic deterogeneous panels. *Journal of the American Statistical Association* 94:621-634.

[26] Pesaran, M. H., Ullah, A., Yamagata, T. (2008). A Bias-adjusted test of error cross section dependence. *The Econometrics Journal* 11:105-127.

[27] Rao, M. R. (1971). Cluster Analysis and Mathematical Programming. *Journal of the American Statistical Association* 66:622-626.

[28] Rota, G. (1964). The number of partitions of a set. *American Mathematical Monthly* 71:498-504.

[29] Sarafidis, V., Wansbeek, T. (2010). Cross-sectional dependence in panel data analysis. Forthcoming in *Econometric Reviews*.

[30] Sealey, C. W., Lindley, J. T. (1977). Inputs, outputs, and theory of production cost at depository financial institutions. *Journal of Finance* 32:1251-1266.

[31] Selim, S. Z., Ismail, M. A. (1984). K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:81-87.

[32] Shao, Q., Wu, Y. (2005). A consistent procedure for determining the number of clusters in regression clustering. *Journal of Statistical Planning and Inference* 135:461-476.

[33] Vahid, F. (1999). Partial pooling: a possible answer to pool or not to pool. In: Engle, R., White, H., ed. *Cointegration, Causality and Forecasting: Festschrift in Honor of Clive W. J. Granger*. Oxford: Oxford University Press.

[34] Wan, S. J., Wong, S. K. M., Prusinkiewicz, P. (1988). An algorithm for multidimensional data clustering. *ACM Transactions on Mathematical Software* 14:153-162.

[35] Yitzhaki, S. (1996). On using linear regressions in welfare economics. *Journal of Business and Economic Statistics* 14:478-486.

Appendix

A Proof of Theorem 2

A1. Overparameterised case: $\Omega_0 < \Omega < \xi$.

Write

$$\begin{aligned} & F_N\left(\Pi_{\Omega}^{(N)}\right) - F_N\left(\Pi_{\Omega_0}^{(N)}\right) \\ = \; & N\log\left[1+\frac{RSS(\Omega)-RSS(\Omega_0)}{RSS(\Omega_0)}\right] + \left[f\left(\Omega\right) - f\left(\Omega_0\right)\right]\theta_N, \\ = \; & N\left(\frac{RSS(\Omega)-RSS(\Omega_0)}{RSS(\Omega_0)} + o\left(\frac{RSS(\Omega)-RSS(\Omega_0)}{RSS(\Omega_0)}\right)\right) + \left[f\left(\Omega\right) - f\left(\Omega_0\right)\right]\theta_N. \end{aligned}$$

We need to show that $F_N\left(\Pi_{\Omega}^{(N)}\right) - F_N\left(\Pi_{\Omega_0}^{(N)}\right) > 0$ a.s. for large N. We know $\left[f\left(\Omega\right) - f\left(\Omega_0\right)\right] > 0$ and, under the conditions of the theorem, θ_N grows faster than $\log\log N$. Further, as $N \to \infty$, $RSS(\Omega_0)/N$ is bounded away from 0 and ∞ almost surely (see, for example, Lemma 2.1 Bai et al. (1999)). Thus the result follows if we can show $RSS(\Omega) - RSS(\Omega_0) = O(\log\log N)$.

We have

$$\begin{aligned} & RSS\left(\Omega\right) - RSS\left(\Omega_0\right) \\ = \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{X}_{C_{\Omega\omega}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_j^0} - \widetilde{X}_{C_j^0}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right\} \\ \geq \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_j^0} - \widetilde{X}_{C_{\Omega\omega}\cap C_j^0}\widehat{\boldsymbol{\beta}}_{\omega|j}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_j^0} - \widetilde{X}_{C_j^0}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right\} \\ = \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\mathbf{Y}}'_{C_{\Omega\omega}\cap C_j^0}\left(I - P_{\widetilde{X}_{C_{\Omega\omega}\cap C_j^0}}\right)\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_j^0} - \sum_{j=1}^{\Omega_0}\widetilde{\mathbf{Y}}'_{C_j^0}\left(I - P_{\widetilde{X}_{C_j^0}}\right)\widetilde{\mathbf{Y}}_{C_j^0}\right\} \\ = \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{\Omega\omega}\cap C_j^0}\left(I - P_{\widetilde{X}_{C_{\Omega\omega}\cap C_j^0}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_j^0} - \sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_j^0}\left(I - P_{\widetilde{X}_{C_j^0}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_j^0}\right\} \\ = \; & T^{-1}\left\{\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_j^0}P_{\widetilde{X}_{C_j^0}}\widetilde{\boldsymbol{\varepsilon}}_{C_j^0} - \sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{\Omega\omega}\cap C_j^0}P_{\widetilde{X}_{C_{\Omega\omega}\cap C_j^0}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_j^0}\right\} \\ = \; & T^{-1}\left\{\sum_{j=1}^{\Omega_0}\boldsymbol{\varepsilon}'_{C_j^0}P_{\widetilde{X}_{C_j^0}}\boldsymbol{\varepsilon}_{C_j^0} - \sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\boldsymbol{\varepsilon}'_{C_{\Omega\omega}\cap C_j^0}P_{\widetilde{X}_{C_{\Omega\omega}\cap C_j^0}}\boldsymbol{\varepsilon}_{C_{\Omega\omega}\cap C_j^0}\right\}, \end{aligned} \tag{28}$$

where the last line follows from the idempotent nature of the matrices $Q_{C_j^0}$ and $Q_{C_{\Omega\omega}\cap C_j^0}$;

$$Q'_{C_j^0}P_{\widetilde{X}_{C_j^0}}Q_{C_j^0} = P_{\widetilde{X}_{C_j^0}}. \tag{29}$$

Under the conditions of the theorem, using (19), we have

$$\boldsymbol{\varepsilon}'_{C_{\Omega\omega}\cap C_j^0} P_{\widetilde{X}_{C_{\Omega\omega}\cap C_j^0}} \boldsymbol{\varepsilon}_{C_{\Omega\omega}\cap C_j^0} = O\left(\log\log N_{\omega|j}\right) = O\left(\log\log N\right) \text{ a.s.}, \tag{30}$$

where $N_{\omega|j} = \left|C_{\Omega\omega} \cap C_j^0\right|$. Thus it follows that $F_N\left(\Pi_\Omega\right) - F_N\left(\Pi_0\right) > 0$ a.s. for N large enough.

A2. Underparameterised case: $\Omega < \Omega_0$.

Again we want to show that for N large enough, $F_N\left(\Pi_\Omega\right) - F_N\left(\Pi_0\right) > 0$ a.s. In this case, $\left(f\left(\Omega\right) - f\left(\Omega_0\right)\right) < 0$ and by assumption, $\lim_{N\to\infty} N^{-1}\theta_N = 0$. The result will follow if we show that $N\log\left(RSS(\Omega)/RSS(\Omega_0)\right)$ is positive and of order N.

The following lemma is necessary for our proof.

Lemma 3 *Suppose that Assumption A.4 holds true. Then, for any possible partition Π_Ω with $\Omega < \Omega_0$, there exist $C_{\Omega\omega} \in \Pi_\Omega$ and $C_{\omega_1}^0, C_{\omega_2}^0 \in \Pi_0$ such that*

$$\left|C_{\Omega\omega} \cap C_{\omega_1}^0\right| > c_0 N \text{ and } \left|C_{\Omega\omega} \cap C_{\omega_2}^0\right| > c_0 N \text{ for any } \omega \text{ and } N \text{ large enough,} \tag{31}$$

where c_0 is a fixed constant.

Proof. See Shao and Wu (2005), Lemma 3.1. ■

From Lemma 3, for any partition $\Pi_\Omega = \{C_{\Omega 1}, ..., C_{\Omega\Omega}\}$, there exists one cluster in Π_Ω, say $C_{\Omega 1}$, and two distinct true clusters C_1^0 and C_2^0, such that

$$c_0 N < \left|C_{\Omega 1} \cap C_1^0\right| < N \text{ and } c_0 N < \left|C_{\Omega 1} \cap C_2^0\right| < N, \tag{32}$$

for N large enough. Denote the family of subsets $\left\{C_{\Omega\omega} \cap C_j^0 : j = 1, ..., \Omega_0,\ \omega = 1, ..., \Omega\right\} - \left\{C_{\Omega 1} \cap C_1^0,\ C_{\Omega 1} \cap C_2^0\right\}$ by $\mathcal{L}_{\overline{12}}$. Then

$$\begin{aligned}
& RSS(\Omega) - RSS(\Omega_0) \\
= \ & T^{-1}\left(\sum_{\omega=1}^{\Omega}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{X}_{C_{\Omega\omega}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_j^0} - \widetilde{X}_{C_j^0}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right) \\
= \ & T^{-1}\left(\left\|\widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_1^0} - \widetilde{X}_{C_{\Omega 1}\cap C_1^0}\widehat{\boldsymbol{\beta}}_{\Omega 1}\right\|^2 + \left\|\widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_2^0} - \widetilde{X}_{C_{\Omega 1}\cap C_2^0}\widehat{\boldsymbol{\beta}}_{\Omega 1}\right\|^2\right) + \\
& T^{-1}\left(\sum_{\mathcal{L}_{\overline{12}}}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_j^0} - \widetilde{X}_{C_{\Omega\omega}\cap C_j^0}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_j^0} - \widetilde{X}_{C_j^0}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right).
\end{aligned} \tag{33}$$

Let $\widetilde{X}_{11} = \widetilde{X}_{C_{\Omega 1} \cap C_1^0}$, $\widetilde{X}_{11a} = \left(\widetilde{X}_{11}' \ \mathbf{0}_{K \times \left| C_{\Omega 1} \cap C_2^0 \right|} \right)'$, $\widetilde{X}_{12} = \widetilde{X}_{C_{\Omega 1} \cap C_2^0}$,

$$\widetilde{\mathbf{Y}}_{012} = \begin{pmatrix} \widetilde{\mathbf{Y}}_{C_{\Omega 1} \cap C_1^0} \\ \widetilde{\mathbf{Y}}_{C_{\Omega 1} \cap C_2^0} \end{pmatrix}, \ \widetilde{X}_{012} = \begin{pmatrix} \widetilde{X}_{11} \\ \widetilde{X}_{12} \end{pmatrix}, \ \widetilde{\boldsymbol{\varepsilon}}_{012} = \begin{pmatrix} \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega 1} \cap C_1^0} \\ \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega 1} \cap C_2^0} \end{pmatrix}. \tag{34}$$

Hence

$$\begin{aligned} & RSS(\Omega) - RSS(\Omega_0) \\ \geq \ & T^{-1} \left(\left\| \widetilde{\mathbf{Y}}_{012} - \widetilde{X}_{012} \widehat{\boldsymbol{\beta}}_{012} \right\|^2 + \sum_{\mathcal{L}_{\overline{12}}} \left\| \widetilde{\mathbf{Y}}_{C_{\Omega\omega} \cap C_{0,j}} - \widetilde{X}_{C_{\Omega\omega} \cap C_{0,j}} \widehat{\boldsymbol{\beta}}_{\omega|j} \right\|^2 \right) \\ & - T^{-1} \sum_{j=1}^{\Omega_0} \widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}' \left(I - P_{\widetilde{X}_{C_{0,j}}} \right) \widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}, \end{aligned} \tag{35}$$

where $\widehat{\boldsymbol{\beta}}_{012}$ is the least squares estimate of $\boldsymbol{\beta}$ based on $\left(\widetilde{\mathbf{Y}}_{012}, \widetilde{X}_{012} \right)$. Since $\widetilde{\mathbf{Y}}_{012} = \widetilde{X}_{012} \boldsymbol{\beta}_{02} + \widetilde{X}_{11a} \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right) + \widetilde{\boldsymbol{\varepsilon}}_{012}$, we have that

$$\begin{aligned} & RSS(\Omega) - RSS(\Omega_0) \\ \geq \ & T^{-1} \left(\widetilde{\mathbf{Y}}_{012}' \left(I - P_{\widetilde{X}_{012}} \right) \widetilde{\mathbf{Y}}_{012} + \sum_{\mathcal{L}_{\overline{12}}} \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega} \cap C_j^0}' \left(I - P_{\widetilde{X}_{C_{\Omega\omega} \cap C_j^0}} \right) \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega} \cap C_j^0} \right) \\ & - T^{-1} \sum_{j=1}^{\Omega_0} \widetilde{\boldsymbol{\varepsilon}}_{C_j^0}' \left(I - P_{\widetilde{X}_{C_j^0}} \right) \widetilde{\boldsymbol{\varepsilon}}_{C_j^0} \\ = \ & T^{-1} \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right)' \widetilde{X}_{11}' \left[I - \widetilde{X}_{11} \left(\widetilde{X}_{11}' \widetilde{X}_{11} + \widetilde{X}_{12}' \widetilde{X}_{12} \right)^{-1} \widetilde{X}_{11}' \right] \widetilde{X}_{11} \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right) \\ & + T^{-1} 2 \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right)' \widetilde{X}_{11a}' \left(I - P_{\widetilde{X}_{012}} \right) \widetilde{\boldsymbol{\varepsilon}}_{012} + T^{-1} \widetilde{\boldsymbol{\varepsilon}}_{012}' \left(I - P_{\widetilde{X}_{012}} \right) \widetilde{\boldsymbol{\varepsilon}}_{012} + \\ & T^{-1} \sum_{\mathcal{L}_{\overline{12}}} \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega} \cap C_j^0}' \left(I - P_{\widetilde{X}_{C_{\Omega\omega} \cap C_j^0}} \right) \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega} \cap C_j^0} - T^{-1} \sum_{j=1}^{\Omega_0} \widetilde{\boldsymbol{\varepsilon}}_{C_j^0}' \left(I - P_{\widetilde{X}_{C_j^0}} \right) \widetilde{\boldsymbol{\varepsilon}}_{C_j^0} \\ = \ & T^{-1} \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right)' \left[\left(\widetilde{X}_{11}' \widetilde{X}_{11} \right)^{-1} + \left(\widetilde{X}_{12}' \widetilde{X}_{12} \right)^{-1} \right]^{-1} \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right) + \\ & T^{-1} 2 \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right)' \widetilde{X}_{11a}' \left(I - P_{\widetilde{X}_{012}} \right) \widetilde{\boldsymbol{\varepsilon}}_{012} - T^{-1} \widetilde{\boldsymbol{\varepsilon}}_{012}' P_{\widetilde{X}_{012}} \widetilde{\boldsymbol{\varepsilon}}_{012} \\ & - T^{-1} \sum_{\mathcal{L}_{\overline{12}}} \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega} \cap C_j^0}' P_{\widetilde{X}_{C_{\Omega\omega} \cap C_j^0}} \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega} \cap C_j^0} + T^{-1} \sum_{j=1}^{\Omega_0} \widetilde{\boldsymbol{\varepsilon}}_{C_j^0}' P_{\widetilde{X}_{C_j^0}} \widetilde{\boldsymbol{\varepsilon}}_{C_j^0}, \end{aligned} \tag{36}$$

using the algebraic identity $(A + B)^{-1} = A^{-1} - A^{-1} \left(A^{-1} + B^{-1} \right)^{-1} A^{-1}$, where A and B are non-singular matrices.

Since we have assumed $\left| \boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right| > 0$, given Assumption A.4 we have

$$\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right)' \left[\left(\widetilde{X}_{11}' \widetilde{X}_{11} \right)^{-1} + \left(\widetilde{X}_{12}' \widetilde{X}_{12} \right)^{-1} \right]^{-1} \left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right) \geq c_0 N \left| \boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02} \right|.$$

Using (17), (19) and the Cauchy-Schwartz inequality we see that the other terms in the above lower bound are of smaller order in N. As $RSS(\Omega_0)/N$ is bounded away from 0 and ∞ almost surely, we have, for N large enough,

$$N \log \left(1 + \frac{RSS(\Omega) - RSS(\Omega_0)}{RSS(\Omega_0)} \right) > N \log(1 + K),$$

for some positive K, and the result follows.

Table A1. Simulation results for $\Omega_0 = 1$.

	$K=1$				$K=4$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
MIC_1								
$\Omega=1$	**.015**	**.000**	**.015**	**.000**	**.000**	**.000**	**.000**	**.000**
$\Omega=2$	.934	.023	.934	.023	.000	.000	.000	.000
$\Omega=3$	.051	.977	.051	.977	1.00	1.00	1.00	1.00
MIC_2								
$\Omega=1$	**.021**	**.032**	**.021**	**.032**	**.533**	**.698**	**.533**	**.698**
$\Omega=2$	.979	.968	.979	.968	.369	.302	.369	.302
$\Omega=3$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_3								
$\Omega=1$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega=2$	.000	.000	.000	.000	.027	.000	.027	.000
$\Omega=3$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega=1$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega=2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=3$	.000	.000	.000	.000	.000	.000	.000	.000

Table A2. Simulation results for $\Omega_0 = 2$.

	$K=1$				$K=4$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
MIC_1								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	**.109**	**.000**	**.231**	**.000**	**.172**	**.000**	**.210**	**.000**
$\Omega=3$	.281	.000	.401	.069	.222	.011	.381	.033
$\Omega=4$	.610	1.00	.368	.931	.606	.989	.409	.967
MIC_2								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	**.196**	**.299**	**.279**	**.412**	**.258**	**.323**	**.396**	**.431**
$\Omega=3$	.383	.589	.332	.587	.495	.376	.501	.569
$\Omega=4$	.421	.112	.396	.001	.247	.301	.103	.000
MIC_3								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega=3$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=4$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega=1$	.069	.009	.000	.000	.086	.000	.000	.000
$\Omega=2$	**.931**	**.991**	**1.00**	**1.00**	**.914**	**1.00**	**1.00**	**1.00**
$\Omega=3$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=4$	.000	.000	.000	.000	.000	.000	.000	.000

Table A3. Simulation results for $\Omega_0 = 3$.

	$K=1$				$K=4$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
MIC_1								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.000	.000	.000	.000	.000	.000	.000	.028
$\Omega=3$	**.541**	**.496**	**.603**	**.568**	**.026**	**.000**	**.049**	**.000**
$\Omega=4$	.459	.504	.397	.432	.439	.000	.491	.000
$\Omega=5$	.000	.002	.000	.000	.535	1.00	.460	1.00
MIC_2								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=3$	**.671**	**.727**	**.715**	**.768**	**.811**	**.310**	**.863**	**.391**
$\Omega=4$	.329	.273	.285	.232	.189	.479	.137	.491
$\Omega=5$	.000	.000	.000	.000	.000	.211	.000	.118
MIC_3								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=3$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega=4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=5$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.088	.027	.000	.000	.151	.049	.000	.000
$\Omega=3$	**.912**	**.973**	**1.00**	**1.00**	**.849**	**.951**	**1.00**	**1.00**
$\Omega=4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=5$	.000	.000	.000	.000	.000	.000	.000	.000

Table A4. Finite sample properties of estimators.

	$K=1$, $\Omega_0=2$, $\overline{\beta}=0.85$, $\boldsymbol{\beta}=(1,.5)'$				$K=1$, $\Omega_0=3$, $\overline{\beta}=0.475$, $\boldsymbol{\beta}=(1,.5,\text{-}.25)'$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
$\widehat{\boldsymbol{\beta}}_p$	.684	.671	.685	.677	-.026	-.031	-.024	-.037
	(.013)	(.008)	(.009)	(.005)	(.008)	(.005)	(.005)	(.003)
$\widehat{\boldsymbol{\beta}}_1$	1.03	1.01	1.02	1.01	1.03	1.02	1.01	1.00
	(.028)	(.018)	(.016)	(.009)	(.034)	(.017)	(.022)	(.009)
$\widehat{\boldsymbol{\beta}}_2$	.505	.502	.502	.501	.501	.501	.501	.502
	(.019)	(.011)	(.010)	(.007)	(.020)	(.010)	(.014)	(.008)
$\widehat{\boldsymbol{\beta}}_3$					-.247	-.251	-.251	-.250
					(.009)	(.005)	(.005)	(.004)